Journal of Biomedical Informatics 56 (2015) 69–79

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

A weighted rule based method for predicting malignancy of pulmonary nodules by nodule characteristics

Aydın Kaya *, Ahmet Burak Can

Hacettepe University, Computer Engineering Department, 06800 Ankara, Turkey

ARTICLE INFO

Article history:
Received 28 December 2014
Revised 18 April 2015
Accepted 15 May 2015
Available online 22 May 2015

Keywords:
Nodule characteristic
Ensemble classifier
Rule based classification
Unbalanced data

ABSTRACT

Predicting malignancy of solitary pulmonary nodules from computer tomography scans is a difficult and important problem in the diagnosis of lung cancer. This paper investigates the contribution of nodule characteristics in the prediction of malignancy. Using data from Lung Image Database Consortium (LIDC) database, we propose a weighted rule based classification approach for predicting malignancy of pulmonary nodules. LIDC database contains CT scans of nodules and information about nodule characteristics evaluated by multiple annotators. In the first step of our method, votes for nodule characteristics are obtained from ensemble classifiers by using image features. In the second step, votes and rules obtained from radiologist evaluations are used by a weighted rule based method to predict malignancy. The rule based method is constructed by using radiologist evaluations on previous cases. Correlations between malignancy and other nodule characteristics and agreement ratio of radiologists are considered in rule evaluation. To handle the unbalanced nature of LIDC, ensemble classifiers and data balancing methods are used. The proposed approach is compared with the classification methods trained on image features. Classification accuracy, specificity and sensitivity of classifiers are measured. The experimental results show that using nodule characteristics for malignancy prediction can improve classification results.

1. Introduction

Lung cancer is one of the leading causes of cancer related deaths worldwide [29]. In the diagnosis of lung cancer, detection of solitary pulmonary nodules is a challenging process for radiologists. A solitary pulmonary nodule is a lung lesion with a diameter of about 2–30 mm and indistinct boundaries. These nodules are generally found fortuitously on tomography scans [1]. With the advances in screening technologies, detection rate of nodules are increased. Computer Aided Diagnosis (CAD) Systems are developed to help radiologists as a second reader. There are two main concerns on CAD Systems; detection and classification of nodules. The challenge in evaluation of a patient's nodule is to determine whether it's benign or malignant. Diagnoses made by radiologists are highly subjective and can be significantly different depending on the level of radiologists' experience.

One of the difficulties in this research area is to find well organized and consistent data. Publicly accessible Lung Image Database Consortium (LIDC) database [2] provides researchers with CT images, nodule region of interests and nodule characteristics as radiographic descriptors. In this database, all cases are evaluated by four radiologists. Each radiologist gives his/her estimations for the boundaries and characteristic ratings of nodules. As Zinovev et al. [9] states, radiologist anonymity and lack of ground truth in LIDC database are challenges; however the database provides the opportunity to build different computer aided diagnosis methods.

In this study, a weighted rule based method for malignancy prediction on pulmonary nodules is presented. The first goal of the study is to show the usefulness of nodule characteristics in malignancy prediction. Separate datasets are defined for each nodule characteristic by applying majority voting on LIDC data. Since most datasets are highly unbalanced, data balancing methods are applied on the datasets. In addition, features are ranked for each nodule characteristic. Subsequent to feature ranking, different feature set sizes are determined for each characteristic by using average ranks and success rate ratios. Since ensemble classifiers are used as a tool to handle unbalanced datasets [31,34,24,25], nodule characteristics are classified with ensemble classifiers. A separate ensemble classifier is built for each nodule characteristic. In the ensemble classification, LDA [35], SVM [28], kNN [37] Adaboost [38], and Random Forest [36] classifiers are tested as the base classifier. Outputs of ensembles are used as inputs for a

* Corresponding author. Tel.: +90 312 297 75 00/157; fax: +90 312 297 75 02.
E-mail addresses: aydinkaya@cs.hacettepe.edu.tr, aydinkaya83@gmail.com (A. Kaya), abc@cs.hacettepe.edu.tr (A.B. Can).

http://dx.doi.org/10.1016/j.jbi.2015.05.011
1532-0464/

weighted rule based method, where the rules are constructed from radiologists' evaluations on nodule characteristics for previous cases. Thus, the expert opinion in LIDC dataset is utilized in the rule based method to predict malignancy. The correlation between malignancy and other nodule characteristics are analyzed to understand the importance of each characteristic in malignancy determination. Furthermore, evaluations of radiologists for the same nodule are analyzed to figure out level of agreement among experts in relation to nodule characteristics. The general schema of the proposed work is shown in Fig. 1.

In later sections, we first give some information about related work. Then, in the methodology section, we give details on LIDC dataset, extracted image features, dataset balancing, feature extraction and feature size determination, and classification steps. After presenting experiments, we discuss the results of our work and future plans.

2. Related work

Most studies on lung nodule detection and classification use only image features to classify lung nodules. With databases like LIDC and NELSON Trial, different challenges have emerged. Handling multiple annotator assessments, providing objective evaluation, predicting other nodule characteristics besides malignancy, and using semantic characteristics to improve malignancy prediction are some of these challenges. We give brief information about some studies which use the LIDC dataset and deal with nodule characteristics. Detailed literature information can be found in surveys by Suzuki [3], Sluimer et al. [4], El-Baz et al. [5].

Zhao et al. [6] propose a CAD system for estimating malignancy of nodules. In this system, ensembles of linear classifiers with feature subsets are constructed. Majority voting is applied on classifier outputs to find probability of malignancy. Jabon et al. [8] develop a content and semantic based image retrieval system that takes CT images as input and retrieves similar images by using image features and semantic characteristics. Euclidean and cosine similarity measures are used in this study. Median voting is used to find the summarized rating for a nodule with multiple annotators. Zinovev et al. [7] propose a method which uses an ensemble decision tree classifier with active learning to predict nodule characteristics. Active learning uses radiologists' agreements on characteristic ratings and predicts the nodules on which radiologists do not agree. The results obtained are better compared to those obtained using

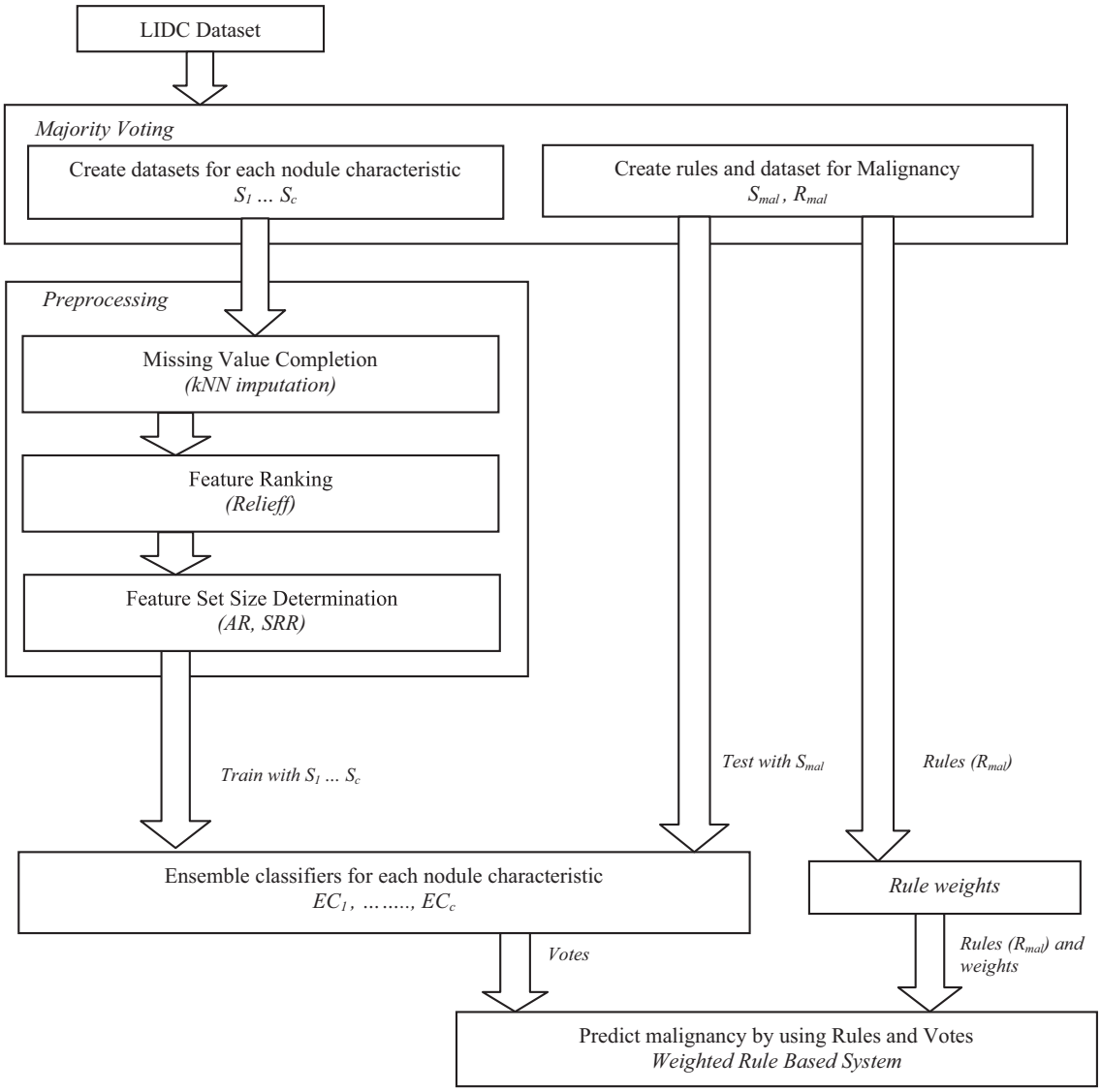
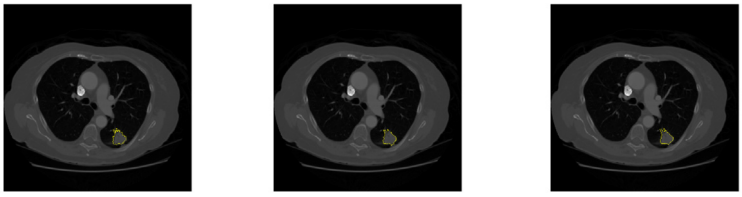


Fig. 1. General schema of the proposed method.



Reading Session 1	
Sublety:	*5*
Internal Str:	*1*
Calcification:	*6*
Sphericity:	*3*
Margin:	*3*
Lobulation:	*3*
Spiculation:	*4*
Texture:	*5*
Malignancy:	*5*

Reading Session 2	
Sublety:	*5*
Internal Str:	*1*
Calcification:	*6*
Sphericity:	*3*
Margin:	*4*
Lobulation:	*4*
Spiculation:	*5*
Texture:	*5*
Malignancy:	*5*

Reading Session 3	
Sublety:	*5*
Internal Str:	*1*
Calcification:	*6*
Sphericity:	*5*
Margin:	*4*
Lobulation:	*1*
Spiculation:	*5*
Texture:	*4*
Malignancy:	*4*

Fig. 2. CT slice of a LIDC case. Same nodule is evaluated by three radiologists. Border of nodules and panel opinions on characteristics are given separately.

single classifiers on nodule characteristics. Zinovev et al. [9] propose a system for predicting nodule characteristics by ensemble of probabilistic classifiers based on belief decision trees and ADABoost learning. They combine the ADABoost approach with belief decision trees to handle the uncertainty of the diagnosis process caused by multiple annotation. They make several adaptations to these methods to tackle the unbalanced data problem of LIDC. They compare the results with those of the single belief decision trees method. In both studies, Zinonev et al. use image features to predict nodule characteristics. Vinay et al. [12] also use image features and concatenate radiographic descriptors to classify malignancy. They examine the response of different classifier types to classify the unbalanced datasets. Instance based, function based, rule based, decision tree based and ensemble based classifiers are compared. Vinay et al. [32,33] extend their work later with different kinds of ensemble classifiers. They found that ensemble classifiers perform better than other classifiers when processing unbalanced data. Lee et al. [11] develop a two-step feature selection method and an ensemble classifier for computer-aided diagnosis of pulmonary nodules. They use random subspace and genetic algorithm methods to select feature subspaces, and combine the results with ensemble methods. However, they do not use LIDC dataset and provide their own dataset of for the experiments. Clinical data and morphological characteristics (such as calcification, cavitation, margin, etc.) and image features are used as features. Their method needs both image features and characteristic evaluations of radiologists to evaluate malignancy of a test case.

Some of the studies try to address issues caused by inter-observer variability. Li et al. [10] use nodule characteristics to improve the objectivity of malignancy prediction and to handle the inter-observer variability. They propose a three layer artificial neural network to predict malignancy with semantic characteristics of pulmonary nodules. They evaluate the subtlety, texture, margin and geometric characteristics using image features. Using these characteristics and different radiologist assessments, they improve the objectivity of the prediction and predict malignancy rating. They compare the results of their approach to those of radiologists by means of Kappa statistics. They find that their method produces more objective results.

Horsthemke et al. [13] use outlines of nodules and nodule characteristics (spiculation, margin, lobulation and sphericity) in LIDC database. They attempt to eliminate the observer disagreement by combining observers' evaluation of nodule areas using probability maps. Image features are extracted from these areas and classification methods are used to predict the combined opinion of the annotators.

3. Method

3.1. LIDC dataset

National Cancer Institute has formed a demand in 2001 for a lung CT image data warehouse which can be accessed via the Internet under the heading of "*Lung Image Database Resource for Imaging Research*". For this purpose, with the efforts of five academic institutions (Cornell University, University of California, University of Chicago, University of Iowa and University of Michigan) a consensus achieved image database, "Lung Image Database Consortium", has developed [2].

For the evaluation of the images, radiologists are assigned from four different institutions. Evaluation phase is divided into two steps: *blinded* (first evaluation of case) and *unblinded* (evaluation after taking consideration of other readers) reads. Same steps occur in two phases: radiologist should detect all possible nodules in a CT scan and provide information about the structure of nodules. A suspicious region that has size smaller than 3 mm is also marked as suspicious but not evaluated. After *blinded* reading session is finished, all information of different institutions are gathered and distributed again. Thus, each radiologist takes into consideration the evaluation of other radiologists and then can edit his/her decisions. Only the *unblinded* reading phase results are added to the database. Results of a nodule evaluation case are shown in Fig. 2.

Each case folder contains DICOM images of the related CT scan and an XML file, which contains panel opinions on nodule characteristics and marked nodule areas of each reading session. Nodule characteristics are calcification, lobulation, subtlety, sphericity, internal structure, spiculation, margin, texture, and malignancy. Table 1 gives a description of nodule characteristics (radiographic descriptors) from Dasovich et al. [14]. Each of them are rated 1 to 5 or 6 by radiologists. This dataset is a valuable resource for researchers who develop CAD systems. Final LIDC dataset contains 1010 cases and information about all nodule matches of different radiologists [15].

3.2. Features

155 image features are calculated for each nodule sample. There are shape, size, and texture-based features. Some features are

Table 1
Nodule Characteristic descriptions and ratings.

Nodule characteristic	Description	Ratings
Calcification	Calcification appearance in the nodule	1. Popcorn 2. Laminated 3. Solid 4. Non-central 5. Central 6. Absent
Internal structure	Expected internal composition of the nodule	1. Soft tissue 2. Fluid 3. Fat 4. Air
Lobulation	Whether lobular shape is apparent from margin or not	1. Marked 2. 3. 4. 5. None
Malignancy	Likelihood of malignancy	1. Highly unlikely 2. Moderately unlikely 3. Indeterminate 4. Moderately suspicious 5. Highly suspicious
Margin	How well defined the margins are	1. Poorly defined 2. 3. 4. 5. Sharp
Sphericity	Dimensional shape in terms of roundness	1. Linear 2. 3. Ovoid 4. 5. Round
Spiculation	Degree of exhibition of spicules	1. Marked 2. 3. 4. 5. None
Sublety	Contrast between nodule and surroundings	1. Extremely subtle 2. Moderately subtle 3. Fairly subtle 4. Moderately obvious 5. Obvious
Texture	Internal density of nodule	1. Non-solid 2. 3. Part Solid 4. 5. Solid

extracted from the largest nodule slice, some from all nodule slices and nodules surrounding area. Nodule surrounding area is obtained by dilating nodule area with a six pixel diameter disk structure element.

High and low degree Zernike moments [16] are obtained from the largest nodule area. Zernike moments are insensitive to rotation but highly sensitive to translation. We placed nodule region of interest to 128×128 square area and then calculated Zernike moments. We used Zernike moment parameters as in Tahmasbi et al.'s [17] work. 32 low-order and 32 high-order Zernike moments included in datasets.

Eccentricity, solidity, circularity, aspect ratio, area of bounding box, standard deviation, and gray level co-occurrence matrix features and Haralick texture features [18] are extracted from the largest area of a nodule, average of all nodule slices, and nodule surrounding areas. Haralick features are contrast, correlation, energy, sum of squares, sum of average, sum of variance, sum of entropy, difference variance, difference entropy, and information measure of correlation. Other gray level co-occurence matrix based features are autocorrelation, contrast, correlation, cluster prominence, cluster shade, dissimilarity, energy, entropy, homogeneity, maximum probability [19], inverse difference normalized and inverse difference moment normalized [20].

3.3. Determining characteristic datasets

In LIDC dataset, it is hard to directly obtain ground truth data for nodule characteristics since annotators may not agree on characteristic ratings. Hence majority voting is used for determining ground truth data. This method assumes all annotators are equally good on subject. Agreement of more than fifty percent of annotators on a decision accepted as the ground truth. For binary labeled data, general definition for majority voting for R annotators can be expressed as in Eq. (1) [21]. y_i is class label for ith sample and y_i^j is the label of annotator j for sample i.

$$y_i = \begin{cases} 1, & (\frac{1}{R})\sum_{j=1}^{R} y_i^j > 0.5 \\ 0, & (\frac{1}{R})\sum_{j=1}^{R} y_i^j < 0.5 \end{cases} \quad (1)$$

Using majority voting, we created an individual dataset for each characteristic. To add a nodule sample in a characteristic dataset, agreement of at least three radiologists is expected on this characteristic. This provides us separate datasets for each characteristic with different amount of nodules. In Table 2, a sample nodule with four radiologist evaluation is shown. If an agreement occurs on a characteristic rating of the nodule, it is included in the characteristic's dataset. Thus this sample is added to all characteristic datasets except sphericity and lobulation datasets. Ochs et al. [30] investigate impact of different level of agreement among radiologists on LIDC database. In relation to this research, we used agreement of 3 or 4 radiologists as a result of majority voting. The latest release of LIDC database contains 1010 cases and 2635 distinct nodules. With our ground truth we selected 438 distinct nodules on malignancy, and 1402 different samples according to radiologist roi definitions.

General description of LIDC nodule dataset S_{LIDC} is given as in Eq. (2). In the equation, x_i is ith image feature, m is the number of features, y_j is the rating for jth *nodule* characteristic, c is the number of nodule characteristics, and n is the number of nodules.

$$S_{\text{LIDC}} = \{x_1, \ldots, x_m, y_1, \ldots y_c\}^{1..n} \in \mathbb{R}^m \quad (2)$$

Table 2
Four radiologists' evaluations on a nodule. Bold values on each column added to related characteristic dataset.

	Subtlety	Calcification	Sphericity	Margin	Lobulation	Spiculation	Texture	Malignancy
Radiologist A	**5**	**6**	3	3	3	**5**	**5**	**5**
Radiologist B	**5**	**6**	3	**4**	4	**5**	**5**	**5**
Radiologist C	**5**	**6**	5	**4**	1	**5**	4	4
Radiologist D	4	3	4	**4**	3	**5**	**5**	**5**

Table 3
AR and SRR ranks for feature set sizes. Bold values are the best scores for each row.

Characteristic	Feature set size										
	Rank	3	4	5	6	7	8	9	10	20	40	60
Calcification	AR	92.65	92.50	87.10	75.85	69.40	58.90	**56.35**	132.55	132.55	145.90	129.25
	SRR	1.0821	1.0830	1.0851	1.0898	1.0919	1.0947	**1.0972**	1.0672	1.0672	1.0630	1.0676
Lobulation	AR	77.80	68.35	33.70	33.10	**18.10**	146.80	115.75	126.70	145.45	179.65	179.35
	SRR	1.1535	1.1564	1.2151	1.2172	**1.2421**	1.0082	1.0596	1.0495	1.0155	0.9913	0.9916
Margin	AR	128.40	119.20	109.80	105.90	156.40	154.60	148.45	109.80	152.35	104.50	**89.65**
	SRR	1.0643	1.0680	1.0735	1.0749	1.0543	1.0549	1.0578	1.0735	1.0554	1.0761	**1.0825**
Sphericity	AR	**14.05**	15.40	18.55	109.80	91.30	92.65	141.40	150.40	130.00	142.00	142.30
	SRR	**1.8496**	1.8380	1.8287	1.0322	1.0486	1.0501	0.9569	0.9424	0.9636	0.9625	0.9627
Spiculation	AR	167.40	163.30	163.30	**29.65**	36.10	31.15	31.45	30.25	91.15	119.95	163.30
	SRR	1.0270	1.0290	1.0290	**1.1686**	1.1651	1.1674	1.1671	1.1680	1.0676	1.0426	1.0290
Sublety	AR	12.40	**4.90**	80.35	62.65	106.5	96.25	75.25	67.15	77.35	192.85	187.75
	SRR	1.5351	**1.5503**	1.1909	1.2138	1.1268	1.1465	1.1770	1.1880	1.1705	0.8726	0.8775
Texture	AR	40.90	39.55	**7.15**	28.45	47.65	122.50	104.20	80.65	87.70	201.85	194.95
	SRR	1.2643	1.2655	**1.3357**	1.2846	1.2499	1.0490	1.0781	1.1180	1.1150	0.9020	0.9245

Input:

t: feature vector of the sample.

S: dataset for the characteristic with n samples, m features, $S = \{x_1, \ldots, x_m, y\}^{1..n}$

m_c: maximum class (rating) label for the characteristic.

F: sorted features by Relieff method.

f_S: feature set size for the binary classifiers obtained from AR and SRR methods.

Begin:

For i=1 to m_c

Create dataset for rating i (class i vs others) $S_i = \{x_1, \ldots, x_m, y_i = \{0,1\}\}^{1..n}$

Balance data, $S'_i = \begin{cases} oversample, & size(S_i(y=1)) < \frac{size(S)}{2m_c} \\ S_i(y=1), & otherwise \end{cases}$

Train the binary learner T_i = Train(S'_i, F_i (1 … f_S))

End

Initialize the vote vector v, $\forall i$: $v(i) = 0$

For i=1 to m_c

If classify (t, T_i) = 1 then

$v(i) = v(i) + 1$

End

End

Output:

v : vote vector of the sample for characteristic ratings.

Fig. 3. Algorithm of ensemble classification for a nodule characteristic of a sample.

3.4. Missing value completion

In datasets, some samples have missing values in consequence of small nodule size. We use k-nearest neighbor imputation [22] method for missing value completion. This method finds nearest k neighbors of a sample by using sample's non missing values. Missing value is imputed by averaging neighbor's values of the related feature. Different methods like expectation maximization, regression, single value decomposition, etc. can also be used for this problem [22].

Table 4
A vote matrix with votes for each characteristic.

Vote matrix	Votes for ratings (1 to 6)					
	(1)	(2)	(3)	(4)	(5)	(6)
Subtlety	0	0	0	0	1	–
Calcification	0	0	0	0	0	1
Sphericity	0	0	1	1	0	–
Margin	0	0	0	1	0	–
Lobulation	0	1	1	1	1	–
Spiculation	1	1	0	0	0	–
Texture	0	0	0	0	1	–

Table 5
Sample rule table for malignancy with nodule characteristics.

Sub.	Cal.	Sph.	Mar.	Lob.	Spi.	Tex.	Mal.
5	3	4	5	2	1	5	1
5	3	5	5	1	1	5	1
5	6	4	2	4	5	5	1
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
4	6	4	5	1	1	5	3
4	6	5	5	1	1	5	3
4	6	4	4	2	2	5	3

3.5. Determining feature sets

Number of selected features can be changed arbitrarily, or defined according to performance cost requirements. Selected features are intended to facilitate class discrimination especially on underexpressed classes. This procedure decreases computational cost of classification while preserving the information.

For each characteristic datasets, we first analyzed the importance of features by using Relieff method [27]. Relieff [27] method sorts features from the most important to less important by assigning weights. Then we used Brazdil and Sores' [26] ranking methods for choosing a suitable feature set size for each characteristic dataset. Ranking methods are average ranks (AR), success rate ratios (SRR) and significant wins (SW). In this study, we use only AR and SRR for evaluation. These ranking methods are generally used for determining which classification algorithm has better performance on a problem. However, we use these methods to determine the appropriate feature set size for the base classifier.

In AR method, classification algorithms are ordered by measured error rates. The best algorithm with the lowest mean error rate is defined as the best. Let r_j^i be the rank of features set size j on bootstrap i and n be the number of bootstraps. Average rank is calculated as in Eq. (3) The lowest rank determines the best feature set size.

Table 6
Three rules from malignancy dataset, which belong to the same nodule. These rules are obtained from three radiologists that agreed on malignancy rating. Agreement ratio is given in a separate column.

Characteristic	$Rule_{i-1}$		$Rule_i$		$Rule_{i+1}$	
	Rating	Agreement	Rating	Agreement	Rating	Agreement
Calcification	6	2/3	6	2/3	3	1/3
Sphericity	3	2/3	4	1/3	3	2/3
Lobulation	5	1/3	1	2/3	1	2/3
Texture	5	3/3	5	3/3	5	3/3
Margin	5	1/3	3	2/3	3	2/3
Subtlety	2	3/3	2	3/3	2	3/3
Spiculation	1	1/3	3	1/3	4	1/3
Malignancy	5	–	5	–	5	–

$$r_j = \left(\sum_i r_j^i\right) \Big/ n \tag{3}$$

In SRR method, success rate between pair of algorithms are evaluated. Let ER_j^i be measured error rate of feature set size j on bootstrap i. In Eq. (4) advantage on feature set size j to k on bootstrap i is calculated as $SRR_{j,k}^i$.

$$SRR_{j,k}^i = (1 - ER_j^i)/(1 - ER_k^i) \tag{4}$$

After calculating advantages, *pairwise mean success rate ratio* $SRR_{j,k}$ for each pair of feature sizes j and k are calculated as in Eq. (5). n stands for bootstrap number. This ratio stated as an estimation of general advantage/disadvantage for pairs.

$$SRR_{j,k} = \left(\sum_i SRR_{j,k}^i\right) \Big/ n \tag{5}$$

Finally, *overall mean success rate ratio* SRR_j is calculated for each feature set size k as in Eq. (6). m is the number of values for feature set size.

$$SRR_j = \left(\sum_k SRR_{j,k}\right) \Big/ (m - 1) \tag{6}$$

We applied these methods on the same classification algorithm with different feature set sizes (3,4,5,6,7,8,9,10,20,40,60). For each characteristic, 10% of dataset used for test data and remaining for training data. Training and testing data are selected with bootstrap method and this procedure is repeated for 50 times. Each bootstrap data are tested on SVM classifiers which are trained with different feature set sizes. In Table 3, mean AR and SRR values for each feature size and characteristics are given. The most meaningful feature set size is implied with bold.

3.6. Dataset balancing

Most of the nodule characteristic datasets are highly unbalanced. For an example, in subtlety dataset, 71% of all nodules are marked 5, 19% are marked 4, 9% are marked 3, and remaining nodules are marked 1 or 2. Similar situation is also observed in the other nodule characteristic datasets. To get rid of inadequate expression of small sample class problem, balancing methods are used.

Data are oversampled with Synthetic Minority Over-sampling Technique (SMOTE) [23], if a rating value is below a ratio. SMOTE method artificially generates synthetic samples, rather than by over-sampling with replacement. Depending on the amount of over-sampling required, minority class is over-sampled by taking k-nearest neighbors of the sample. After a sample and a nearest neighbor are chosen, differences between their feature vectors are calculated. Multiplying these differences with a random number between 0 and 1, and adding this vector to the sample vector generates a new sample that lies on a random point along the segment between these two samples in consideration [23]. It's a simple and effective method to create synthetic samples and prevents over-fitting problems like oversampling with replacement (replicate the samples) approach.

3.7. Ensemble classifiers

After preprocessing the dataset, ensemble classifiers are built based on the algorithm given in Fig. 3. The algorithm takes feature vector of a sample, characteristic dataset, maximum class label, features sorted by Relieff method [27], and feature set sizes as inputs; and generates a vote vector as output. In the first loop, characteristic datasets and ensemble classifiers are built. For each rating of a nodule characteristic, a training dataset is built and a separate binary classifier is trained. For example, five training sets are formed for rating values (1–5) of subtlety characteristic. A feature set is selected for each classifier based on the feature rankings of Relieff method and the feature set size is determined in the preprocessing step. If dataset is unbalanced, data balancing procedures are also applied as explained in the previous section. Each rating value and other rating values (all remaining classes) are considered as two separate classes. In this way, five binary classifiers are built corresponding to five rating values of subtlety characteristic.

During the testing, test samples are given to these separate classifiers. In this study, all binary classifiers are assumed to have the same weight. If a binary classifier produces positive result for a sample, relevant rating vote increased by one. A final vote matrix for a nodule sample is shown in Table 4. In this table, the votes with '1' value means that related binary classifier produced positive result for the sample. In the experiments, 2-class LDA, SVM, kNN Adaboost, and Random Forest classifiers are used as the base classifier.

3.8. Weighted rule based method

Subsequent to voting phase, a weighted rule based method is used to predict malignancy rating of nodules. The rule set used in this method obtained from the malignancy dataset. Radiologists' evaluations on nodule characteristics for previous cases are used to construct rules. A sample rule set is shown in Table 5. In the rule set, the nodules that at least three radiologists agreed on malignancy values are included.

Malignancy dataset is created using majority voting on malignancy rating but an agreement over other nodule ratings is not expected for this dataset. Since an agreement is not expected on nodule ratings, our approach defines coefficients to measure importance of nodule ratings for each rule. Different coefficients are assigned for characteristics of each rule, based on agreement ratio of characteristics. For example, in Table 6, three rules from

Input:

t: feature vector for sample.

C: number of characteristics.

N: number of rules for predicting malignancy.

R: rule matrix.

mr: malignancy values for each rule as a list.

α : correlation coefficient.

β : agreement ratio.

F: sorted features by Relieff method.

f_S: feature set size for base classifier obtained from AR and SRR methods.

Begin:

Initialize the vote matrix, V

For i = 1 to C (for each characteristic)

S_i: dataset for i^{th} characteristic

m_c^i: maximum rating value for i^{th} characteristic

Get votes from ensemble classifier, $V(i)=ensembleClassifier(t,S_i, m_c^i, F, f_S)$

End

For j=1 to r (for each rule)

Find weighted vote total for each rule, $P(j,1) = \sum_{i=1}^{C} e^{\alpha_i} V\left(i, R(j,i)\right) e^{\beta_j^i}$

Add true malignancy rating for each rule, $P(j,2) = mr(j)$

End

Sort P in descending order by weighted vote totals, $P' = sort(P(,1))$

For i = 1 to 5 (for each malignancy rating)

Find the probability of rating i, $p(i)$ = (count of rating i in the first N elements of P') / N

End

Find the rating value with maximum probability, $m = \{i: \max_{i=1,\ldots,5} p(i)\}$

End

Output:

p: malignancy probability of the sample.

m: malignancy rating with the highest probability

Fig. 4. Algorithm of the weighted rule based method.

Table 7
Truth vs classification results of subtlety for a 5-class random forest with unbalanced dataset.

	Classification result					Sample size	Sample ratio
	1	**2**	**3**	**4**	**5**		
Truth							
1	**3**	0	6	3	3	15	<0.01
2	0	**0**	7	1	1	9	<0.01
3	0	0	**77**	41	13	131	~0.09
4	0	0	22	**162**	87	271	~0.19
5	0	0	7	39	**984**	1030	~0.71

Bold values are correctly classified samples.

Table 8
Truth vs classification results of subtlety for a 5-class random forest with balanced dataset.

	Classification result					Sample size	Sample ratio
	1	2	3	4	5		
Truth							
1	**8**	0	2	3	2	15	<0.01
2	1	**1**	4	2	1	9	<0.01
3	0	0	**100**	22	9	131	~0.09
4	1	1	59	**145**	65	271	~0.19
5	1	0	24	52	**953**	1030	~0.71

Bold values are correctly classified samples.

Table 9
Truth vs. the votes for subtlety ratings for random forest ensemble classifier.

	The votes					Sample size	Sample ratio
	1	2	3	4	5		
Truth							
1	**8**	8	5	4	4	15	<0.01
2	3	**4**	4	3	3	9	<0.01
3	43	42	**103**	43	44	131	~0.09
4	91	91	94	**138**	97	271	~0.19
5	54	51	85	80	**981**	1030	~0.71

Bold values are correctly assigned samples.

the rule set which belong to a nodule are shown. $rule_{i-1}$, $rule_i$, and $rule_{i+1}$ are three radiologists' observations and their agreement on malignancy rating is 5. Two of them agreed on calcification rating and their agreement ratio on calcification are 2/3; and other's ratio is 1/3. All radiologists agreed on texture/subtlety rating and their ratio are 3/3.

Additionally, our approach uses correlation analysis to discover relationships among each characteristics and malignancy. The correlation coefficient only measures the degree of linear association between two characteristics. Strongly correlated characteristics may dominate over results in calculations. Hence this coefficient is used with exponential function to smooth its effect as in Eq. (7).

$$P(j) = \left\{ \sum_{i=1}^{C} e^{\alpha_i} V(i, R(j,i)) e^{\beta_j^i}, mr(j) \right\} \tag{7}$$

In Eq. (7), C is number of characteristics in dataset (we use 7 characteristics), V is a $C \times 6$ voting matrix, where each row represents votes for a nodule characteristic as in Table 4. R is the $r \times C$ rule table. j is the rule number in R. α_i is correlation coefficient of characteristic i on malignancy rating. β_j^i is agreement rating of ith characteristic of jth rule. $P(j)$ holds two values: $P(j,1)$ as weighted vote total of jth rule and $P(j,2)$ *as* the malignancy rating (mr) of jth rule. Afterwards, P is sorted in descending order by weighted vote totals. Malignancy probability of test sample is determined by occurrences of malignancy ratings in first N rule.

Algorithm in Fig. 4 shows weighted rule based classification of a sample nodule. In the first *for* loop, the algorithm obtains vote vectors of nodule characteristics and generates vote matrix by using the algorithm in Fig. 3. In the second *for* loop, weighted vote totals (P) for each row in the rule matrix are calculated by using Eq. (7). After sorting P, probabilities of malignancy ratings are calculated by counting occurrences in the first N elements of P. Finally, the class label with maximum probability is given as predicted malignancy rating.

4. Results

4.1. Ensemble classifier results

Single classifiers are not generally very effective comparing to ensemble ones on unbalanced datasets [34]. To observe this issue, we conducted experiments with single and ensemble versions of our best performing Random Forest classifier. Confusion matrices of single Random Forest classifier trained for subtlety characteristic with unbalanced and unbalanced dataset are shown in Tables 7 and 8 respectively. For rating 1 and 2, sample sizes are lower than 1% of all data. For the dominant rating 5, this ratio is over 70%. As it can be observed from Table 7, misclassification ratios of underexpressed classes by single classifier with unbalanced dataset are high. With balanced dataset, misclassification ratio is reduced with the same classifier.

Ensemble classifiers produce a vector result (votes for characteristic ratings) instead of a single classification result. This vector voting mechanism increases representation of underexpressed ratings. In Table 9, the votes for subtlety characteristics are shown. Comparing to Table 8, ensemble Random Forest classifier performed a little better than single Random Forest classifier except class value of 4. In Table 9, bold values on the diagonal are the votes achieved for the related rating (true class). Other values on the table are different ratings that take votes besides the related rating. We call this situation an ambiguity. For example, in Table 4, subtlety ensemble produces 1 vote only for rating 5. However, lobulation ensemble gives 1 for all but rating 1. This is a highly ambiguous result for a classification phase. Therefore, lobulation characteristic has lower distinctive effect for predicting malignancy, but subtlety, calcification, and texture has more distinctive values for the sample in Table 4. Such ambiguities are solved by the weighted rule based method and a higher classification performance is obtained.

4.2. Weighted rule based method results

Leave-one-out procedure is used for method evaluation. A single sample is chosen from the original malignancy dataset as the validation data, and its corresponding rule is removed from the rule set. The remaining sample's rules (not features) are used as the training data. Each sample has a case id and nodule id. Selected sample is also excluded from nodule characteristic datasets if datasets have a nodule with the same case-nodule id. This procedure is repeated until each sample in the dataset is used once as the validation data.

For the algorithm in Fig. 4, different rule sizes are tested to find the optimum value for N. Scales are chosen between 10 and 100. According to Figs. 5 and 6, the highest score with the lowest rule set size is around scale 30. Hence optimal rule set size is set to 30 in the experiments.

The performance of our weighted rule based method and different kind of classifiers are compared: instance based k-nearest neighbor (kNN); ensemble based Adaboost; function based Support Vector Machines (SVM), linear discriminant classifier (LDA); tree and ensemble based Random Forest (RF) and naïve bayes classifier. These classifiers are trained over malignancy dataset using image features. Similar to our approach, k-nearest neighbor imputation is applied on malignancy dataset for completion of

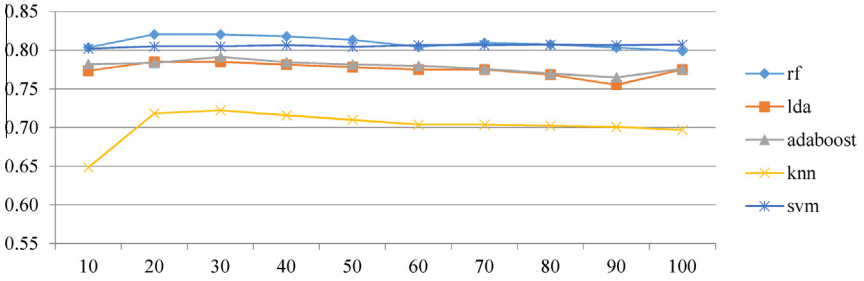


Fig. 5. Rule set size effect on classification accuracy on 5 class experiment.

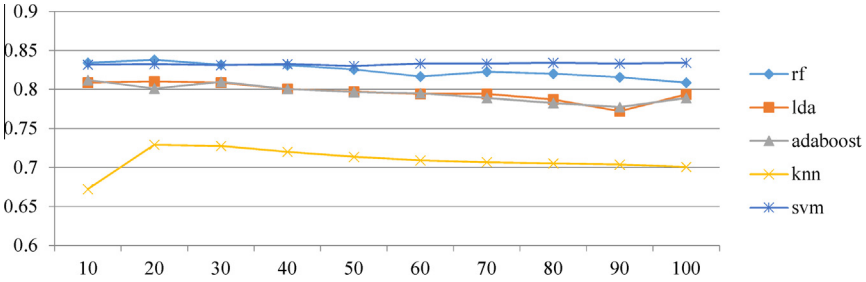


Fig. 6. Rule set size effect on classification accuracy on 3 class experiment.

missing values before testing. Leave-one-out procedure is used to test all classifiers.

Performances of classifiers are evaluated according to classification accuracy, sensitivity, specificity and statistical significance on classification error. Malignancy has 5 ratings: Highly unlikely (1), moderately unlikely (2), indeterminate (3), moderately suspicious (4), and highly suspicious (5). In the first experiment, classifiers are tested over 5 class ratings on malignancy dataset. Classification accuracy, sensitivity and specificity values are shown in Table 10 for 5 class malignancy rule set. Table 11 shows *p*-values of significance test on classification error for single classifiers vs. our method with different base classifiers.

In the second experiment, 3 class malignancy dataset is formed by grouping ratings (1, 2) as unlikely, 3 as indeterminate, (4, 5) as suspicious. In this experiment the scatter among similar ratings is reduced. In Table 12, classification accuracy, sensitivity and specificity measures are given. Table 13 shows p-values of significance test on classification error for 3 class classifiers.

5. Discussion

In the 5 class experiment, according to Table 10, RF and SVM ensemble classifiers obtained 82.52% and 80.60% classification accuracy respectively. Although LDA and Adaboost ensemble classifiers have better performance than most of the single classifiers with 78.53% and 79.14% classification accuracy respectively, their results are lower than single RF classifier. RF single classifier obtained the highest classification accuracy among all single classifiers with score of 80.40%. In the context of specificity, our best performing RF and SVM ensemble classifiers and single RF classifier have the highest scores with 94.74%, 94.20% and 94.21% respectively. All the single and ensemble classifiers have more than 90% specificity except single LDA classifier. Sensitivity of our ensembles is lower than single RF, kNN, NB, and Adaboost classifiers. However, sensitivity of all tested classifiers is very low comparing to specificity values. This shows that these classifiers perform better in identifying the negative values than identifying positive values. According to the significance analysis results in Table 11, in 5 class experiment, there is significant difference between RF ensemble and single classifiers ($p < 0.05$) except RF single classifier ($p = 0.168$). SVM ensemble is another well performing method but there is no significant difference of SVM ensemble with single SVM ($p = 0.131$) and single RF ($p = 0.862$) classifiers.

In the 3 class experiment, the scatter between similar ratings is reduced by grouping similar ratings. According to Table 12, RF ensemble classifier obtained the best classification accuracy with 84.89% and SVM ensemble classifier obtained the second best score with 83.36%. Other ensemble classifiers LDA, Adaboost, and kNN have lower performance. The most notable increase in classification accuracy is obtained on LDA single classifier with 16% improvement. Single RF classifier obtained 81.51% classification accuracy, which is minor increase comparing to 5 class experiment. In the context of specificity and sensitivity, our best performing RF and SVM ensemble classifiers have the highest scores (92.09%, 83.11%) and (91.17%, 82.59%) respectively. In the 3 class experiment, classification accuracy and sensitivity are improved for all single and ensemble classifiers. Identifying positive and negative sample probability are close in this experiment. According to the significance analysis results on Table 13, there is significant difference between RF ensemble and all single classifiers ($p < 0.05$). While kNN ensemble results are also significant ($p < 0.05$), this significance is negative since classification results of kNN ensemble is lower comparing to all single classifiers.

Table 10
Classification accuracy, sensitivity, and specificity of tested methods (5 ratings).

Method	CA	Sens	Spec
LDA	0.6610	0.3832	0.8886
Adaboost	0.7254	0.6737	0.9311
Naive Bayes	0.7250	0.6700	0.9325
kNN	0.7665	0.6314	0.9350
SVM	0.7828	0.5384	0.9345
RF	0.8040	0.5797	0.9421
Ensemble SVM	0.8067	0.5575	0.9420
Ensemble RF	0.8252	0.5586	0.9474
Ensemble LDA	0.7853	0.5450	0.9349
Ensemble Adaboost	0.7914	0.5433	0.9360
Ensemble kNN	0.7224	0.4479	0.9128

Table 11
Significance test (*p*-values) on classification error for ensemble methods vs single methods (5 ratings).

	E (RF)	*E* (SVM)	*E* (Adaboost)	*E* (kNN)	*E* (LDA)
LDA	0.000	0.017	0.000	0.000	0.000
Adaboost	0.000	0.000	0.000	0.864	0.000
Naive Bayes	0.000	0.000	0.000	0.882	0.000
kNN	0.000	0.012	0.125	0.010	0.250
SVM	0.001	0.131	0.592	0.000	0.877
RF	0.168	0.862	0.423	0.000	0.237

Table 12
Classification accuracy, sensitivity and specificity of methods (3 ratings).

Method	CA	Sens	Spec
LDA	0.8200	0.8001	0.9006
Adaboost	0.8135	0.8265	0.9055
Naive Bayes	0.8100	0.8250	0.9060
kNN	0.7959	0.8003	0.8930
SVM	0.8005	0.8108	0.9021
RF	0.8151	0.8161	0.9035
Ensemble SVM	0.8336	0.8259	0.9117
Ensemble RF	0.8489	0.8311	0.9209
Ensemble LDA	0.8090	0.8082	0.9001
Ensemble Adaboost	0.8098	0.8000	0.8999
Ensemble kNN	0.7278	0.6701	0.8590

Table 13
Significance test (p-values) on classification error for ensemble methods vs single methods (3 ratings).

	E (RF)	E (SVM)	E (Adaboost)	E (kNN)	E (LDA)
LDA	0.047	0.359	0.503	0.000	0.470
Adaboost	0.016	0.178	0.809	0.000	0.769
Naive Bayes	0.008	0.115	0.990	0.000	0.948
kNN	0.001	0.013	0.372	0.000	0.401
SVM	0.001	0.029	0.549	0.000	0.584
RF	0.021	0.214	0.729	0.000	0.690

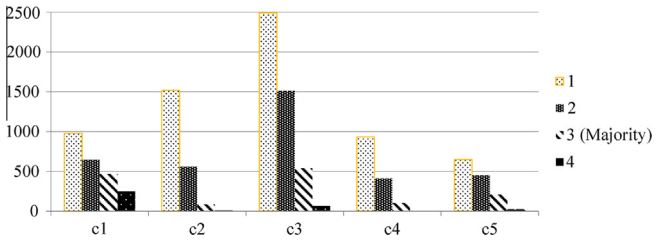


Fig. 7. Distribution of malignancy ratings in different ground truths (agreement of at least 1–4 radiologists).

When comparing 5 class and 3 class experiment results, classification accuracy and sensitivity generally improved around average of 4% and 25% respectively. Specificity reduced around average of 3%. When single and ensemble classifiers are compared, excluding kNN ensemble classifier, ensemble of a base classifier have better performance than single classifier of the same type in 5 class experiment. In 3 class experiment, kNN and LDA ensembles have lower performance comparing to single classifier of their types. Fig. 7 shows the distribution of malignancy ratings in different ground truths (agreement of at least 1 to 4 radiologists). Since we used majority voting in our experiments, agreement of at least 3 radiologists is accepted as the ground truth. Although a decline is expected in sample size when consensus increases, number of moderate rating samples decrease more comparing to other ratings. This situation shows that radiologists tend to agree on less ambiguous ratings, which are highly unlikely (1), indeterminate (3), and highly suspicious (5). Grouping ratings into 3 classes decreases uncertainty among similar ratings. Therefore, 3-class experiment generally produces better results comparing to 5-class experiment.

Although LIDC dataset is publicly available dataset, comparing studies in the literature is difficult due to varying assumptions. In the literature, Zinovev et al. [7,9] and Vinay et al. [12,32,33] have prominent studies on malignancy prediction using LIDC dataset. Zinovev et al. [7] used median voting on different agreement ratios (at least one/two/three radiologists) and June 2009 release of LIDC database (207 patients with 914 distinct nodules). They proposed several experiments and reported 0.7251 ± 0.18 classification accuracy on average. In another study, Zinovev et al. [9] reported classification accuracy of 0.5900 ± 0.04 on malignancy prediction. In these studies, they use active learning and ensemble classifiers to improve classification accuracy over single classifiers. Vinay et al. [12,32,33] used another 2009 release of LIDC database with 399 cases. First, they evaluated different classifier performances on this dataset [12] and they proposed and compared several ensemble classifier based methods in the subsequent studies [32,33]. They reported 0.7828 ± 0.06 average classification accuracy. However, it is not clear how the ground truth is determined. Making a comparison among these studies is difficult due to varying generalization methods, datasets and ground truth assumptions. Although the comparison is difficult, our results are appreciable within similar studies in the literature.

6. Conclusion

In this paper, usefulness of nodule characteristics in malignancy prediction is studied. Latest release of LIDC database with 1010 case is used. Majority voting on radiologist agreements is used to determine the ground truth data. A rule based method is proposed for malignancy prediction on pulmonary nodules. This method takes a rule set extracted from panel agreements on LIDC dataset. Rules are weighted according to ratio of radiologist agreements and correlation between characteristics and malignancy ratings. Separate datasets are defined for each nodule characteristic. To handle unbalanced nature of the datasets, ensemble classifiers, dataset balancing methods, and class specific feature selection are used. Ensemble classifiers with SVM, Random Forest, LDA, kNN and Adaboost base classifiers are built for each nodule characteristic. Different feature set sizes are determined for each characteristic by using average ranks and success rate ratios methods. Then, outputs (votes) of ensembles are used as inputs for the rule based method.

Experimental results showed that nodule characteristics can be used to improve classification results on malignancy prediction. Our ensemble classifiers with SVM and Random Forest base classifiers have performed better than other types of classifiers in most of the results. In the malignancy dataset, moderate ratings have fewer samples and highly/moderately suspicious and highly/moderately unlikely ratings are close evaluations. These divided assessments may affect the classification results negatively. Grouping malignancy ratings have improved classification performance.

For the future work, we plan to expand the feature set and combine the ensemble classifiers. Different ratios of radiologist agreements can be investigated as the ground truth. In this paper, we tried to predict malignancy by only using semantic characteristic information in the rule based classification phase. Using image features and characteristic information together may improve results further. Combining the information obtained from radiologists' evaluations by using probability maps and median voting can also be studied to find alternative approaches.

Conflict of interest statement

The authors declare that there are no conflict of interest.

Acknowledgments

The authors would like to thank Dr. Murat Dündar at IUPUI, Dr. Ali Seydi Keçeli at Hacettepe University and Berkan Yılmaz at Bilkent University for their valuable opinions and contributions.

References

[1] J.H. Austin, N.L. Müller, P.J. Friedman, D.M. Hansell, D.P. Naidich, M. Remy-Jardin, et al., Glossary of terms for CT of the lungs: recommendations of the Nomenclature Committee of the Fleischner Society, Radiology 200 (2) (1996) 327–331.
[2] S.G. Armato III, G. McLennan, M.F. McNitt-Gray, C.R. Meyer, D. Yankelevitz, D.R. Aberle, et al., Lung image database consortium: developing a resource for the medical imaging research community, Radiology 232 (3) (2004) 739–748.
[3] K. Suzuki, machine learning in computer-aided diagnosis of the thorax and colon in CT: a survey, IEICE Trans. Inform. Syst. 96 (4) (2013) 772–783.
[4] I. Sluimer, A. Schilham, M. Prokop, B. van Ginneken, Computer analysis of computed tomography scans of the lung: a survey, IEEE Trans. Med. Imaging 25 (4) (2006) 385–405.
[5] A. El-Baz, G.M. Beache, G. Gimel'farb, K. Suzuki, K. Okada, A. Elnakib, et al., Computer-aided diagnosis systems for lung cancer: challenges and methodologies, Int. J. Biomed. Imaging (2013), http://dx.doi.org/10.1155/2013/942353. Article ID 942353:46.
[6] L. Zhao, M.C. Lee, L. Boroczky, V. Vloemans, R. Opfer, Comparison of computer-aided diagnosis performance and radiologist readings on the LIDC pulmonary nodule dataset, in: Medical Imaging. International Society for Optics and Photonics, 2008, pp. 69151I–69151I-8.

[7] D. Zinovev, D. Raicu, J. Furst, S.G. Armato III, Predicting radiological panel opinions using a panel of machine learning classifiers, Algorithms 2 (4) (2009) 1473–1502.
[8] S.A. Jabon, D.S. Raicu, J.D. Furst, Content-based versus semantic-based retrieval: an LIDC case study, in: SPIE Medical Imaging. International Society for Optics and Photonics, 2009; pp. 72631L–72631L-8. http://dx.doi.org/10.1117/12.812877.
[9] D. Zinovev, J. Furst, D. Raicu, Building an Ensemble of Probabilistic Classifiers for Lung Nodule Interpretation, in: 10th International Conference on Machine Learning and Applications and Workshops, 2011, vol. 2, pp. 155–161.
[10] G. Li, H. Kim, J.K. Tan, S. Ishikawa, Y. Hirano, S. Kido, et al., Semantic characteristics prediction of pulmonary nodule using Artificial Neural Networks. Engineering in Medicine and Biology Society (EMBC), 35th Annual International Conference of the IEEE 2013, vol. 3–7, pp. 5465–5468, http://dx.doi.org/10.1109/EMBC.2013.6610786.
[11] M.C. Lee, L. Boroczky, K. Sungur-Stasik, A.D. Cann, A.C. Borczuk, S.M. Kawut, et al., Computer-aided diagnosis of pulmonary nodules using a two-step approach for feature selection and classifier ensemble construction, Artif. Intell. Med. 50 (1) (2010) 43–53.
[12] K. Vinay, A. Rao, G.H. Kumar, Prediction of lung nodule characteristic rating using best classifier model, Int. J. Comput. Appl. 56 (18) (2012) 29–32.
[13] W.H. Horsthemke, D.S. Raicu, J.D. Furst, Predicting LIDC diagnostic characteristics by combining spatial and diagnostic opinions, in: SPIE Medical Imaging. International Society for Optics and Photonics, 2010, pp. 76242Y–76242Y-9.
[14] G.M. Dasovich, R. Kim, D.S. Raicu, J.D. Furst, A model for the relationship between semantic and content based similarity using LIDC, in: SPIE Medical Imaging. International Society for Optics and Photonics, 2010, pp. 762431–762431-10.
[15] The Cancer Imaging Archive Wiki Site: <https://wiki.cancerimagingarchive.net/display/Public/Wiki> (last access 01.07.14).
[16] F. Zernike, Diffraction theory of the cut procedure and its improved form, the phase contrast method, Physica 1 (1934) 689–704.
[17] A. Tahmasbi, F. Saki, S.B. Shokouhi, Classification of benign and malignant masses based on Zernike moments, Comput. Biol. Med. 41 (8) (2011) 726–735.
[18] R.M. Haralick, L.G. Shapiro, Computer and Robot Vision: vol. 1, Addison-Wesley, 1992, p. 459.
[19] L. Soh, C. Tsatsoulis, Texture analysis of SAR sea ice imagery using gray level co-occurrence matrices, IEEE Trans. Geosci. Remote Sens. 37 (2) (1999) 780–795.
[20] D.A. Clausi, An analysis of co-occurrence texture statistics as a function of grey level quantization, Can. J. Remote Sens. 28 (1) (2002) 45–62.
[21] V.C. Raykar, S. Yu, L.H. Zhao, G.H. Valadez, C. Florin, L. Bogoni, et al., Learning from crowds, J. Mach. Learn. Res. 11 (2010) 1297–1322.
[22] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, et al., Missing value estimation methods for DNA microarrays, Bioinformatics 17 (6) (2001) 520–525.
[23] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (1) (2002) 321–357.
[24] P. Yang, W. Liu, BB. Zhou, S. Chawla, AY. Zomaya, Ensemble-based wrapper methods for feature selection and class imbalance learning, in: Advances in Knowledge Discovery and Data Mining, Springer, Berlin, Heidelberg, 2006, pp. 544–555.
[25] H. Guo, H.L. Viktor, Learning from imbalanced data sets with boosting and data generation: the DataBoost-IM approach, SIGKDD Explor. Newsl. 6 (1) (2004) 30–39.
[26] P. Brazdil, C. Soares, A comparison of ranking methods for classification algorithm selection, in: 11th European Conference on Machine Learning, Springer, Berlin, Heidelberg, 2000, pp. 63–75.
[27] Y. Sun, J. Li, Iterative RELIEF for feature weighting, in: Proceedings of the 23rd International Conference on Machine Learning ACM, 2006, pp. 913–920.
[28] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297.
[29] World Health Organization Fact Sheet: <http://www.who.int/mediacentre/factsheets/fs297/en/ (last access 01.07.14).
[30] R. Ochs, H.J. Kim, E. Angel, C. Panknin, M. McNitt-Gray, M. Brown, Forming a reference standard from LIDC data: impact of reader agreement on reported CAD performance, in: Proc. SPIE 6514, Medical Imaging 2007: Computer-Aided Diagnosis, 65142A (30.03.07), http://dx.doi.org/10.1117/12.707916.
[31] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, F. Herrera, A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches, IEEE Trans. Syst. Man. Cybern C 42 (4) (2012) 463–484.
[32] K. Vinay, A. Rao, G.H. Kumar, Classifiers in context: prediction of radiological characteristic ratings for lung nodule malignancy, Int. J. Appl. Inform. Syst. 5 (2) (2013) 14–19.
[33] K. Vinay, A. Rao, G.H. Kumar, Combining Ensemble of Classifiers Using Voting-Based Rule to Predict Radiological Ratings for Lung Nodule Malignancy. Emerging Research in Electronics, Computer Science and Technology, Springer India, 2014, pp. 443–451.
[34] V. Nikulin, G.J. McLachlan, S.K. Ng, Ensemble approach for the classification of imbalanced data, In AI 2009: Advances in Artificial Intelligence, Springer, Berlin, Heidelberg, 2009, pp. 291–300.
[35] R.A. Fisher, The use of multiple measurements in taxonomic problems, Ann. Eugen. 7 (2) (1936) 179–188.
[36] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[37] E. Fix, J.L. Hodges, Discriminatory Analysis, Nonparametric Discrimination: Consistency Properties. Technical Report 4, USAF School of Aviation Medicine, Randolph Field, Texas, 1951.
[38] Y. Freund, R.E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, J. Comput. Syst. Sci. 55 (1) (1997) 119–139.